Subject Centre for Languages, Linguistics and Area Studies
Call for project funding bids

Application form

Title of proposal:	**A Web-as-Corpus approach to populating Wikiversity for teaching about Islam and Muslims in language, linguistics and area studies**
Lead proposer:	Dr Eric Atwell and Prof James Dickins
Institution:	University of Leeds
Address:	School of Computing and School of Modern Languages and Cultures, University of Leeds Leeds LS2 9JT
Phone:	0113 3435430
Fax:	0113 3435468
Email:	E.S.Atwell@leeds.ac.uk and J.Dickins@leeds.ac.uk
Names of additional proposers:	
Total funding requested: (Max £3,000)	£3,000

Provide a brief description (about 100 words) of your project. If your bid is successful, this will be used to announce the project via print and electronic media.

Wikiversity is an online open-source public repository for University-level teaching and learning materials, based on the Wikipedia architecture for “crowd-sourcing”: it relies on volunteers to collaborate by actively contributing their knowledge for the common good. Undoubtedly a wealth of learning resources exists on the WWW, but scattered on individual websites, in a wide range of formats and structures. Individual lecturers prepare online teaching materials to support their own teaching, but few know of Wikiversity, and few have time or inclination to take on the extra step of formally registering and uploading their materials to Wikiversity. This project aims to organise and semi-automate the harvesting of these scattered resources, by adapting Web-as-Corpus techniques from Corpus Linguistics.

It is important that you complete this form in conjunction with the guidelines for applications. No section should be longer than 500 words.

1. Full description of project and proposed outcomes, including relevance of the project to the development of teaching and learning in Islamic Studies in UK higher education

Language and Linguistics research and teaching often involves a Corpus or empirical sample of the language being studied. The WWW is increasingly used as a corpus source, at three levels:

- A web search engine such as Google or Yahoo allows linguists to view the WWW as a corpus, by searching for and extracting empirical evidence of language items directly from the Web

- Scholars can "scout" the Web for suitable text-samples fitting their research needs, and download them to compile a specialised corpus, eg the Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 2006)

- A "robot" can be used to automate the process of harvesting a corpus fitting specified criteria; for example BootCat is given a list of specific "seed terms" and uses these to harvest WWW documents matching some or all of these seed terms, eg (Atwell et al 2009, Atwell et al 2011)

We propose to adapt and combine these three complementary approaches, to harvest Wikipedia-level teaching and learning resources for teaching and learning about Islamic Studies. This will involve:

- Investigation into search strategies and terms to identify candidate teaching materials appropriate to Islamic Studies university courses

- "scouting" for resources: applying these search strategies and terms to find and download a "candidate corpus" of web-based resources which appear potentially useful for Islamic Studies teaching and learning

- Adapting BootCat, the web-as-corpus robot, to automatically harvest potential Islamic Studies teaching materials; for this, we need to identify seed-terms and parameter constraints which enable BootCat to target potentially suitable websites.

In parallel to this three-pronged approach to "harvesting", we need to develop an evaluation procedure, to filter out unsuitable candidate teaching resources. Potential resources must be reviewed by academic teaching staff in Arabic and Middle Eastern Studies, who will use their expertise and judgement to evaluate (i) relevance to their own teaching, and (ii) potential usefulness to the wider community of teachers and learners about Islam and Muslims. Resources judged potentially useful will be followed up: we will contact the authors for agreement to upload to Wikiversity, of course acknowledging the original source; and given agreement, we will upload to Wikiversity, aiming to develop a broad but coherent set of teaching resources.

This outline could be a programme for a large-scale research and development project; but given the limited budget of £3000, we will be restricted to a pilot feasibility study. We will use the funding to hire a Summer Intern: a talented

Computing and Artificial Intelligence student employed over summer 2011 to adapt the three web-as-corpus techniques to our task of harvesting candidate teaching materials. This technical development and corpus harvesting will be supervised and guided by Atwell, leader of the Arabic Computing research group and experienced with web-as-corpus techniques for Corpus Linguistics research, e.g. (Al-Sulaiti and Atwell 2006, Atwell et al 2009, Atwell et al 2011). Our pilot study will focus on a restricted target "application": teaching and learning about Islam and Muslims in the School of Modern Languages and Cultures at Leeds University. The evaluation of candidate resources will focus on potential for reuse in specific taught modules in this School. Dickins is Professor and Head of Department of Arabic and Middle Eastern Studies and Director of the Language, Linguistics and Translation research group within the School of Modern languages and Cultures; Dickins will evaluate candidate teaching materials for potential reuse in his department's teaching, and also forward candidate teaching materials to colleagues.

The main outcomes of the project will be:

- A collection of online reusable teaching materials for Islamic Studies, integrated into Wikiversity, including a subset evaluated as specifically relevant for taught courses at Leeds University
- A methodology and software support for harvesting further resources, reusable by and/or for other universities
- An empirical evaluation of this methodology and its applicability to teaching and learning about Islam and Muslims; we cannot guarantee in advance that our approach will succeed in identifying high-quality reusable teaching resources, but the only way to test this hypothesis is to try it and see.
- Publication of the pilot study results and lessons learnt, in journal paper(s) – for both Language/linguistics and Computing research communities.

2. Potential benefit to the Islamic Studies higher education community, and potential impact on the student learning experience in Islamic Studies

Please describe the potential benefit of your proposed outcomes and deliverables, including the potential take-up by practitioners and the potential impact on the student learning experience.

The outcomes listed above are of themselves self-evidently beneficial:

- The collection of online reusable teaching materials for Islamic Studies can be reused widely, not just at Leeds University
- The methodology and tools for harvesting further resources can also be reused more widely, again not just at Leeds University
- The evaluation of the methodology will enable us (and others) to decide whether to proceed further, or else perhaps to abandon the idea as impractical.

- Whatever the outcome, publication of the pilot study in journal(s) is a benefit.

Other benefits include:

- Enhanced awareness of online resources and the benefits of reusability
- Collaboration between Language and Computing researchers and scholars, to mutual enrichment
- The new perspective on Islamic Studies via Wikiversity online reused resources should add variety, novelty and hopefully inspiration to the student learning experience

3. Innovation

Please state how your project addresses a clearly identified gap in knowledge or resources, and is innovative in the context of learning and teaching in Islamic Studies.

Applying Web-as-corpus harvesting techniques and technology to Islamic Studies is clearly novel. Probably other proposals to this Call will be from individuals or groups aiming to develop teaching resources for their own use; reusability of resources is central to Software Engineering, but less commonplace in teaching about Islam and Muslims.

The “gap” we will fill is not a specific subtopic in Islamic Studies, but a gap in reuse and sharing of learning resources generically.

The project is also innovative in terms of Computing research: to our knowledge, Web-as-Corpus techniques have not been previously adapted to harvesting teaching materials. If successful, this could open up a new area of applied research in Computing and Artificial Intelligence.

4. Dissemination

How will project outcomes be disseminated to a wide audience?

Wikiversity is an established architecture for sharing and dissemination of teaching resources, which will help us to persuade Language, Linguistics and Area Studies scholars that the resources are ready and easy to re-use. The novelty of our approach to “harvesting the web for Islamic Studies” should also help grab attention, and inspire learners and teachers to try our results. The experimental case study should be publishable as computing research, and this should help publicise the Islamic Studies learning resources as a by-product.

5. Monitoring and Evaluation

How will you evaluate the effectiveness of your project?

In our post-project publications, we can report on the number and range of teaching resources harvested, and the range of taught courses serviced by these harvested resources. This evaluates effectiveness in terms of usefulness for teaching and

learning about Islam and Muslims. We can also apply metrics from Machine Learning and Information Retrieval research (precision, recall, etc) to the software tools, such as our adapted version of BootCat. The most important outcome will hopefully be wide reuse of our contributions to Wikiversity, though this can only be judged in the longer term. Another outcome, if our pilot study is a success, would be to seek funding for a larger-scale follow-up research and development project.

6. Costs and Value for money

Outline the costs involved. Provide details of buying out of staff time and other costs. What contribution will the host institution make? How does the project provide value for money?

Given that only £3000 is on offer, we think the best use of this limited sum is to hire a Summer Intern, a talented Computing and Artificial Intelligence student employed over summer 2011, to develop the software and use it to harvest potential resources. Ideally the supervision and evaluation contributions of Atwell and Dickins ought ot be funded as staff buy-out time; but £3000 is barely sufficient to fund the Summer Intern for 10-12 weeks, so Proposers' time will have to be contributed by Leeds University. The Summer Intern will also be provided with use of desk, PC and research computing infrastructure in our AI Lab, to work alongside externally-funded postgrads and postdocs. So, this is excellent value for money: research on the cheap!

7. Timetable

Please provide a timetable of your work which will be used to monitor progress.

April-June 2011: more detailed planning and background research by Atwell and Dickins, including review of other Wikiversity-related projects

July-September 2011: Summer Intern employed for main phase of project implementation, including pilot population of Wikiversity with web-harvested teaching resources

October-December 2011: review or results, writing up for publication in journal(s) or conference(s)

8. References

L Al-Sulaiti and E Atwell. 2006. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, vol. 11, pp. 135-171

E Atwell et al. 2009. Arabic and Arab English in the Arab World. Proc CL2009 International Conference on Corpus Linguistics, Liverpool University, UK

E Atwell et al. 2011. An Artificial Intelligence approach to Arabic and Islamic content on the internet. Proc NITS'2011 National Information Technology Symposium, King Saud University, Saudi Arabia.

Data protection statements

- *I understand that the information I have provided will be stored in an electronic format by the Higher Education Academy and its Subject Centres.*
- *I understand that the information I have provided will be accessible to, and shared by, the Higher Education Academy and its Subject Centres.*
- I understand that my name, job title and department may be shared with my employer for networking, professional development and reporting purposes.

Signature of lead proposer:

Agreement of senior management in the institution(s) involved

Signature of Head of Institution or nominee:

Name and position:

Name of partner institution (if any):
Signature of Head of Institution or nominee:
Name and position:

Please return this form both electronically and in hard copy to:

Sue Nash
Subject Centre for Languages, Linguistics and Area Studies
School of Humanities
University of Southampton
Southampton
SO17 1BJ

Deadline for submission: Friday 18 February 2011